LA-UR-21-20341
Approved for public release; distribution is unlimited.

Title:	From the Starship Enterprise to Los Alamos National Laboratory Artificial Intelligence and Machine Learning in the NSRC
Author(s):	Ali, Alee Rizwan
Intended for:	also to be included in the Lab's NSS magazine Web
Issued:	2021-01-14

From the Starship Enterprise to Los Alamos National Laboratory

Artificial Intelligence and Machine Learning in the NSRC

Rizwan Ali

Director, National Security Research Center

Los Alamos National Laboratory

Artificial Intelligence (AI) conjures up different emotions in different people. Some view it with fear, imagining a malevolent AI similar to Skynet in the *Terminator* movies. Others see something benevolent, such as the character Data in *Star Trek: The Next Generation*. Still others see it as a means to advance the frontiers of science, engineering, and technology to new levels not possible through traditional means. At the National Security Research Center (NSRC or the Center), we see AI as a tool to help us go through the monumental tasks we have in digitizing, cataloging, and searching our collections.

Broadly speaking, AI involves developing computers that mimic human intelligence and cognition. The current state of AI is nowhere near anything we can classify as having sentience similar to Skynet or Data, and it's anyone's guess if machine consciousness will ever happen. Speculation about that possibility might be best left to futurists and science fiction authors.

However, a branch of AI, called Machine Learning, has made significant progress over the past couple of decades. Machine Learning, sometimes written as AI/ML, uses sophisticated, digital neural networks, and algorithms to implement self-learning algorithms to perform specific functions. Many of us use AI/ML on a daily basis without thinking much about it. Digital assistants, such as Siri and Alexa, on our smart devices use AI/ML to learn the types of news we like to read and locations we typically travel to on the weekends, among other preferences in our routines. YouTube uses AI/ML to deliver the most-relevant videos for you based on your

recent watch and search history. Smart thermostats know when we usually come home from work, adjusting the temperature so we arrive at a warm, toasty house in the cold months.

Most of us take these advances for granted and don't think twice about the technological sophistication needed for devices and services to learn our behavior and make accurate predictions about our preferences. One significant limitation with AI/ML systems is that they are very specific in terms of what function they can perform. For example, there is no chance, at least for now, that a Nest Thermostat's AI/ML algorithm could be used to predict stock market trends. An AI/ML system must be developed and taught the specific set of tasks for which it was designed.

Those tasks, however, are quite remarkable, and not possible just a generation ago before digital neural networks and sophisticated algorithms employed by AI/ML began to be developed. This is the reason why the NSRC is exploring this technology to truly revolutionize the way we operate, and more importantly, the way we contribute to our nation's security.

Machine Learning and the LANL mission

Implementation of AI/ML technology in the NSRC is a priority for the Center's leadership team in order to better serve our researchers -- the scientists and engineers at LANL who work in support of national deterrence. The NSRC's collections encompass tens of millions of holdings, making the Center one of the largest research libraries in the United States. We estimate that less than 10 percent of these holdings have been digitized, and less than 10 percent of those digitized holdings have been cataloged. This affects the speed with which the Center is able to provide researchers with the Lab's one-of-a-kind materials that are so vital to their work. Without employing AI/ML technologies in multiple areas of the Center, it is unlikely the NSRC will make significant progress in digitizing and cataloging our collections.

Some may be asking why it's important for us to digitize, catalog, and make searchable this vast collection of nuclear weapons material. The short answer: It saves a lot of time and even more money. The Center's holdings are active collections, not an antiquated, rarely-used archive. Implementing AI/ML is a mission-critical task for the Center's collections to be accessible to researchers. The information contained in the NSRC's collections are used on a daily basis by nuclear weapons physicists, engineers, and production specialists in a number of mission areas related to weapons stockpile stewardship, design development, and pit production. The collections have reports and analyses that save the Lab tens of millions of dollars annually and preclude countless hours in redundant research, studies, and experiments. Furthermore, the majority of these records do not exist elsewhere -- if researchers need them, the Center is the only option to get them.

Modernizing equipment and processes

The NSRC is exploring a variety of systems to find the best places to implement AI/ML technologies, including the process to digitize our vast collections of physical material, automated tasks to capture metadata and catalog the digitized information, and the implementation of a natural language search system.

In just one portion of our holdings, we use an AI/ML system to digitize documents in our microfilm and microfiche collections. These collections contain information relevant to nuclear weapons modeling and simulation, weapons designs, and pit production, all of which are critical to the Lab's stockpile stewardship mission and pit production benchmarks at the Lab. Now, we want to extend our application of AI/ML technologies even further and to broader sets of our collections.

The Center's microfiche and microfilm number in the hundreds of thousands and contain well over 50 million pages of information. Using our current, non-AI/ML-capable equipment, software, and processes, it would take us an estimated 90-some years to digitize the microfiche

collections, and over 2,000 years to digitize our microfilm collection. The absence of AI/ML means significant manpower is required to operate the outdated equipment and perform the cumbersome, manual quality-assurance process required for each digitized page.

Using modern AI/ML-based equipment and software, coupled with improved processes, has a high likelihood of cutting the amount of time to digitize this material to under a decade. The AI/ML systems could potentially decrease the time to digitize the microfiche collection by as much as 80 percent and the microfilm collection by as much 99 percent by automatically detecting individual frames and performing highly sophisticated image corrections to automatically flag the dozen or so images out of several thousand that the AI/ML system was not able to automatically correct. This dramatically reduces the time our archivists spend performing quality-assurance (QA) reviews on the finished digitized products.

But digitizing and performing QA on the documents are just two of several steps where AI/ML can be employed. The ultimate goal is not to just digitize the material, but to present researchers with materials that have been cataloged and are easily searchable using a natural-language search system.

A solution to a 400-year backlog

Currently, cataloging in the Center is a manual, time consuming process where our librarians or archivists upload metadata information into one of our classified digital repositories, the Online Vault or PDMLink. The metadata contains information such as the document's title, date, author(s), report number, organization, abstract, and key words. This process can take anywhere from 10 to 30 minutes per document depending on the complexity of the document and speed of the system that particular day.

If the NSRC continued to catalog its current backlog of 2.4 million digitized documents, it would take over 400 years to complete. Meanwhile, as we begin to increase the rate at which we are

digitizing our physical collections, the total number of digitized documents will continue to grow at a rapid pace. There aren't enough qualified people in the NSRC or anywhere in the Lab to manually catalog such vast collections of digitized documents.

To automate the cataloging process, an AI/ML-based system needs to be implemented. This system would perform a sophisticated optical character recognition process and parse the information in each document into metadata, which could then be cataloged. After the metadata is parsed, it would pass the information to an AI/ML-based, natural-language search system. These are multiple, distinct processes that often involve the partnership of several companies.

The NSRC's documents date back more than 75 years to the inception of the Manhattan Project and contain older typewriter fonts that cannot be searched using industry-standard document viewing software, such as Adobe Reader. The AI/ML system needs to read each document regardless of its format and fonts and then extract the required metadata so it can be cataloged. The metadata extraction system scans the digitized documents and, through an AI/ML process, teaches itself where to find the relevant metadata information.

This information is then passed to another system that uses AI/ML to implement a natural-language search tool. The natural-language search tool uses contextual clues in the way a researcher would phrase a search query to discern the intent of the search, rather than deliver just a simple list of documents that contain the words in the search field.

To illustrate this point, suppose a researcher wanted information about the word "plant." The system would use contextual clues in the full search string to determine if the researcher wanted to know about the biological entity "plant," or a manufacturing "plant," or information about how to "plant" something in the ground. Each meaning would yield vastly different results and the use of an AI/ML-based natural-language search system would deliver only the most relevant results to the researcher.

To address this metadata/search challenge, the NSRC initiated a large-scale AI/ML project to automatically catalog our digitized information and provide a natural language search system to assist researchers to find relevant information. Companies that do this highly specialized type of AI/ML are rare. To find these companies, the NSRC reached into the U.S. Intelligence Community (IC), which has a very similar problem set, namely that it has vast quantities of digital information that need to be cataloged and searched in a rapid and efficient manner.

After a fairly lengthy process, the NSRC found a set of companies that specialize in using AI/ML to extract metadata from digitized documents and use natural-language, AI/ML-based systems to search through materials.

A successful test run

The Lab's Deputy Director for the Weapons Program provided the NSRC funding to test the system in a six-month pilot study on the LANL unclassified network, using unclassified digitized nuclear power plant material. The system's AI/ML system was able to successfully capture the required metadata automatically, to include keywords, and populate the cataloging system. Because this system was also used within the IC, it also passed the Lab's classified analysis without any problems, clearing the path for it to be installed on the Lab's classified network.

The end goal for this AI/ML project, called Titan on the Red, is to extract metadata from various digital data repositories, such as Online Vault, shared drives, and SharePoint, and present researchers with a natural language, AI/ML-based search interface to search through the NSRC's entire digitized collection. Because many of the documents within the NSRC are protected through stringent security and need-to-know protocols, the search system will enforce these protocols and only deliver documents that the researcher had the appropriate approvals to view.

What this means for researchers is that the large backlog of documents that are currently not cataloged and not searchable, will become accessible in a matter of months once the system comes online, rather than in double-digit decades. Plus, the process to search through the NSRC's digital repositories will become dramatically easier than the current process, which requires contacting one of our librarians to have him or her manually search through our collections. Note that at least initially, the search capability will only be available to researchers inside LANL.

For the fiscal year 2021, the Center's goal is to:

- install Titan on the Red on the Lab's classified network,
- start the process to integrate the system into one or more of our digital repositories,
- begin the process of training the system's AI/ML neural network to extract the required metadata from the documents, and
- train the AI/ML search system to identify words and terms specific to the types of weapons program documents we have in our collections.

In the beginning, the system will only be available to the NSRC's staff and a select group of researchers, as we fully test the system. The goal is to make the system available to everyone in the Lab's Weapons Program and others who have a need to access weapons program material.

AI/ML is new, yet proven. LANL needs to embrace this advancement, which is really the only solution to making its one-of-a-kind collections searchable to LANL researchers. Investing in AI/ML saves countless manpower hours and many millions of dollars, while directly contributing to the Lab's mission success and our nation's security.

What is the National Security Research Center?
More than 75 years ago, Los Alamos was a clandestine laboratory charged with creating the world's first atomic weapons to help end World War II.

Today, the success of that scientific achievement endures as the foundation of Los Alamos National Lab's national security mission.The National Security Research Center (NSRC) is an integral part of this mission.

The NSRC curates the reports, films, photographs, lab notebooks, engineering drawings, and more that led to the dawn of the Atomic Age. The collections are among the NSRC's tens of millions of materials, which span the entire history of the nuclear enterprise. These include historical documents and artifacts, classified research, and weapons information. The NSRC also supports the Lab's online classified repositories and provides guidance for publishing materials in the classified realm.

Additionally, the NSRC's specialized team of 40 historians, archivists, librarians, and digitizers partner with the Lab's scientists and engineers as they conduct research.

The NSRC opened its doors in 2019, transitioning from what was a repository of archival materials to a dynamic, vibrant library that researchers access on a daily basis. Already, the nascent NSRC is one of the largest research libraries in the federal government.

"The National Security Research Center is a truly remarkable, one-of-a-kind organization with a vital purpose," said Charlie Nakhleh, the Associate Laboratory Director for Weapons Physics. "Today's geopolitical environment underscores our national security mission – and the NSRC is critical in that mission success. These collections support LANL's future innovations in nuclear weapons stewardship and development."

The NSRC ensures the Lab – and the nation – remains vigilant and ready.

Watch:
Powering Tomorrow's Innovations Today
https://www.youtube.com/watch?v=q-ld1hhx44c
infographics x 3
staff photos


TORK